アノテーションに基づくスポーツ映像要約とナレーション生成

大平 茂輝 † 長尾 確 †
† 名古屋大学エコトピア科学研究機構

1 はじめに

映像コンテンツの検索や要約を目指して、従来から様々な映像解析手法や音声処理手法が提案されている。しかし、それらの事前処理は具体的な応用に対してコストが見合っていなかったり、より高度な応用、すなわち意味的な要約や変換を行う際に十分な情報の精度と粒度が備わっていないことがある。スポーツ映像の要約研究には、ハイライト [1] や効率的な映像処理 [2] に重点を置くものなどがあるが、要約のレベルは様々であること、要約の前処理にかかるコストや効果に大きな差が存在することはあまり議論されていない。

本論文では、半自動的なアノテーションに基づくスポーツ映像要約と、要約映像に対するナレーション生成手法を提案する。また、アノテーションの方法に応じて、1) 機械処理のみによって自動生成する要約、2) 機械処理と人手によるアノテーションを用いる要約、3)2) の処理に加え外部知識を用いる要約の 3 種類に分類し、各々について要約の枠組みと実例を示す。

2 半自動的なスポーツ映像アノテーション

筆者らは、スポーツ映像に特化したアノテーションツール SVA (Sports Video Analyzer/Annotator) を試作している。SVA は、スポーツ映像中のチーム情報、選手情報、位置情報、プレイ内容、コメント、実況や解説音声の発話内容をアノテーションとして生成・編集・関連付けすることを可能にする。本ツールによるアノテーション処理の流れを図 1 に示す。

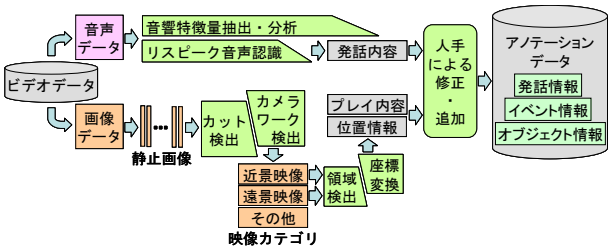

図 1: SVA のアノテーション処理

アノテーションで扱う内容記述は、情景描写の直接の構成要素となるイベント情報、画面中を移動する選手や審判、ボールといったオブジェクト情報、試合経過を実況・解説する発話情報から構成される。イベント情報にはタイムコードとプレイ内容・反則・得点状況、オブジェクト情報にはタイムコードと名称、説明、位置、発話情報にはタイムコードと発話内容が含まれる。たとえばサッカー映像の場合、プレイ内容にはドリブルやパス、シュートなどが含まれ、これらはプロファイルによって定義されたラベルで表される。

Annotation-based Sport Video Summarization and Narration Generation
† Shigeki OHIRA(ohira@esi.nagoya-u.ac.jp)
† Katashi NAGAO(nagao@nuie.nagoya-u.ac.jp)
EcoTopia Science Institute, Nagoya University (†)
Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8603, Japan

機械処理によって得られる映像切替のタイムコード、オブジェクトの位置と出現区間、発話内容には、認識誤りや誤検出による誤った情報が多く残っている。そこで、人手によってそれらの誤った情報を修正すると同時に、人間の目で見れば容易に判断できるプレイ内容や選手名などの情報の追加を行う。

以上の手順を経ることで、イベント情報、オブジェクト情報、発話情報に関するアノテーションデータを作成することができ、それらは XML で記述される。

2.1 映像解析と人手によるメタデータ付与

映像ストリーム中の画像データは、構成要素であるフレーム単位の静止画像ごとにカット/カメラワーク検出が行われ、フィールド近景・フィールド遠景・それ以外の 3 つの映像カテゴリに分類される。近景映像は、事前に入力された選手のユニフォームの色情報を基に、選手とボールの領域検出が行われる。遠景映像は、フィールド上のライン/ボール/選手の認識が行われ、2 次元座標上の位置情報取得に利用される。その他の映像には、観客席やベンチなどの映像が含まれる。

また、SVA では機械処理による選手やボールの位置検出とは別に、人間が視覚的に判断したフィールド上の粗い位置情報の入力を支援する。具体的には、フィールドを 12 のブロックに分割し、テンキーにより映像を視聴しながらリアルタイムに入力を行う (図 2, 3)。

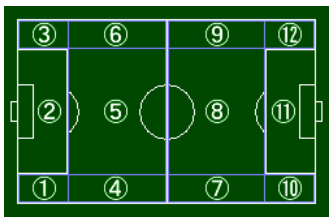

図 2: フィールド分割

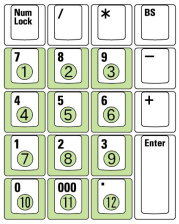

図 3: テンキー配置

放送映像のようにブレのないクリアな映像の場合、単体のカメラ映像からでもフィールド上の位置情報を取得することはそれほど難しくはないかもしれない。しかし、素人の撮影した映像やノイズの多い映像 (たとえば芝の状態が悪い、土のグラウンドに白線等) を対象とする場合、自動処理は極めて困難であり、上記入力支援に基づくメタデータ付与は非常に有効である。

また、このようにして蓄積されるアノテーションデータは、検索・要約のインデックスとしてだけでなく、対象領域を自動検出するための学習に再利用することも可能である。

2.2 実況音声の書き起こし

音声データから抽出される音響特徴量は、観客の盛り上がり区間の検出等に使用される。実況/解説音声の発話内容は、リスピーク方式に基づいてアノテーション作成者が同じ内容を発声し、その音声を認識することにより抽出する。BBC や NHK は、リスピーク方式によるスポーツ中継番組の字幕放送を実現している。

ニュース番組のような読み上げ口調の整った発話においては、高い認識性能が示されているが、スポーツ番組のような歓声などの背景音を含む、複数話者による口語調の発話においては、音声認識が困難である。リスピークに基づく音声認識は、これらスポーツ実況音声認識が抱える問題に対する実用的な解決方法の一つと言える。ただし、本研究では字幕作成が目的ではないため、リスピーカーによる内容の要約は行わず、不要な単語の除去と文法的な誤りの訂正に止めている。

3 スポーツ映像要約

スポーツ映像に対して、選手やボールの位置情報、プレイ内容のラベル、実況の音声情報、音声パワー、といったアノテーションを適切に付与することにより、多様な要約が可能になる。

図 4 は、機械処理のみによって自動生成する要約であり、「観客の盛り上がったシーンを 3 分で視聴したい」という要求に対する要約作成である。この場合に用いるアノテーションデータは、音声パワーを中心とする音響特徴量と、プレイイベントの区間長である。各プレイイベント区間における音響特徴量の平均値を計算し、値の上位から順に 3 分という要約時間を満たすまでプレイイベントを抽出・結合する、というのが基本的な処理である。

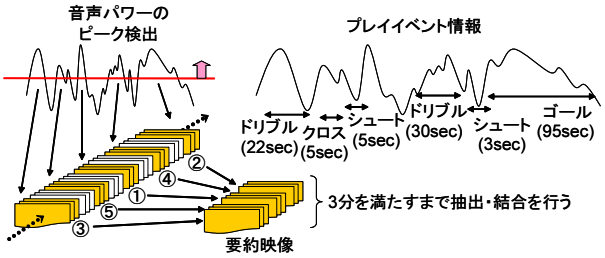

図 4: 「観客の盛り上がったシーンを 3 分で要約」

より滑らかで視聴しやすい要約を作成するための副次的な処理としては、得点シーンの重みを高くしたり、実況音声が途切れないよう始端/終端のタイムコードを調整したり、関連するプレイ同士、たとえばシュートは直前のセンタリングやパスまでを一連のプレイとみなして抽出する、といった処理が有効である。

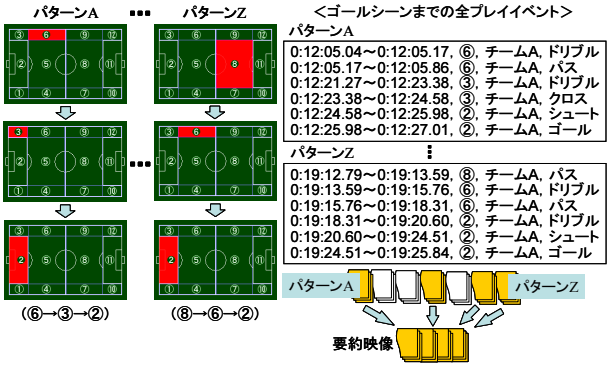

図 5: 「右サイドからの攻撃による得点シーン」

図 5 は、機械処理と人手によるアノテーションを用いる要約であり、「右サイドからの攻撃による得点シーン」という要求に対する要約作成である。使用されるアノテーションデータは、選手とボールの位置情報およびプレイイベント情報である。正確な位置情報を取得できている場合には該当データを用いて、そうでない場合にはテンキーにより入力されたフィールド上の分割ブロック情報から右サイドに対応するプレイイベント区間を抽出し、ゴールシーンまでの全プレイイベントを結合することで要約を得ることができる。

4 要約映像のナレーション生成

機械処理と人手によるアノテーションに加えて外部知識を用いる要約例を示す。試合の内容を時間経過とともに記述したテキストがある場合、各時間の内容記述に含まれるキーワードを用いてそのシーンの重要度を決定し、アノテーションデータ中の対応するプレイイベント区間を重要度順に抽出する、という処理によって要約を作成することができる (図 6)。

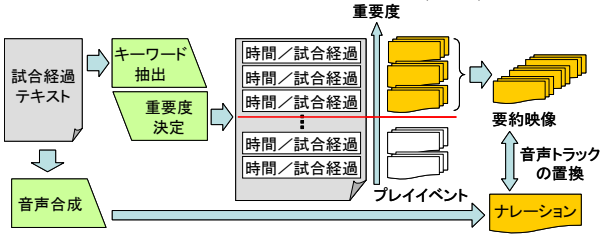

図 6: 試合経過テキストを用いた映像要約

要約映像のナレーション生成手法は 2 通りある。1 つはオリジナルの実況音声を用いる方法で、もう 1 つは合成音声を用いる方法である。前者の方法では、抽出される映像区間の境界と音声区間の境界とは必ずしも一致しないという問題があるため、実況音声が途切れないように音声区間の時間的伸縮や無音区間の除去といった工夫が必要である。後者の方法では、抽出区間に相当する説明文を試合経過テキストから抜き出して音声合成を行い、要約映像の音声ストリームと置換することでナレーションを生成することができる。抽出区間と試合経過テキストがマッチしない場合には、該当区間の時間とプレイイベント情報、位置情報をもとに説明文を生成し、音声合成を行う。

5 まとめと今後の課題

3 種類の映像要約の枠組みと実例を挙げ、自動解析に頼らずに興味深い要約を作成することのできる可能性を示した。また外部知識を効率的に用いることで、内容的に質の高い要約映像のナレーションを生成できることも示した。しかし、あらゆる要約を想定すると問題はそれほど簡単でないことも事実である。このような問題に対して機械処理を適用することも選択肢の一つではあるが、別の解決方法として、同一映像を視聴することの可能な複数ユーザによるアノテーション作業の分散化が考えられる。我々はこれをオンラインアノテーションと呼び、現在検討中である。

参考文献

[1] D. Tjondronegoro, Y-P P. Chen and B Pham, "Integrating Highlights to Play-break Sequences for More Complete Sport Video Summarization," *IEEE Multimedia*, Vol. 11, No. 4, pp. 22-37, 2004.

[2] A. Ekin and M. Tekalp, "Automatic Soccer Video Analysis and Summarization," *IEEE Trans. Image Processing*, Vol. 12, No. 7, pp. 796-807, 2003.